

A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions

S. M. NUSSER, A. L. CARRIQUIRY, K. W. DODD, and W. A. FULLER

The distribution of usual intakes of dietary components is important to individuals formulating food policy and to persons designing nutrition education programs. The usual intake of a dietary component for a person is the long-run average of daily intakes of that component for that person. Because it is impossible to directly observe usual intake for an individual, it is necessary to develop an estimator of the distribution of usual intakes based on a sample of individuals with a small number of daily observations on a subsample of the individuals. Daily intake data for individuals are nonnegative and often very skewed. Also, there is large day-to-day variation relative to the individual-to-individual variation, and the within-individual variance is correlated with the individual means. We suggest a methodology for estimating usual intake distributions that allows for varying degrees of departure from normality and recognizes the measurement error associated with one-day dietary intakes. The estimation method contains four steps. First, the original data are standardized by adjusting for nuisance effects, such as day-of-week and interview sequence. Second, the daily intake data are transformed to normality using a combination of power and grafted polynomial transformations. Third, using a normal components-of-variance model, the distribution of usual intakes is constructed for the transformed data. Finally, a transformation of the normal usual intake distribution to the original scale is defined. The approach is applied to data from the 1985 Continuing Survey of Food Intakes by Individuals and works well for a set of dietary components that are consumed nearly daily and exhibit varying distributional shapes.

KEY WORDS: Continuing Survey of Food Intakes by Individuals; Density estimation; Dietary status; Measurement error models.

1. INTRODUCTION

The U.S. Department of Agriculture (USDA) has been responsible for conducting periodic surveys to estimate food consumption patterns of households and individuals in the United States since 1936. Because dietary intake data from these surveys are used to formulate food assistance programs, consumer education efforts, and food regulatory activities, it is crucial that appropriate methodologies be used in the analysis of these data. An important concept in analyzing food consumption data is *usual intake*, defined as the long-run average of daily intakes of a dietary component by an individual. From a statistical perspective, the usual intake of individual i is defined to be

$$y_i = E\{Y_{ij}|i\},$$

where Y_{ij} is the intake of a dietary component on day j by individual i . This article outlines a methodology to estimate usual intake distributions of dietary components consumed on a nearly daily basis (e.g., nutrients, cholesterol, energy) from 24-hour recall dietary intake data.

To assess usual intake, daily dietary intakes are often collected from individuals for a number of days. If an individual's mean daily intake for a particular dietary component is used as an indication of his or her usual intake, then the variance of the mean intakes contains some within-individual variability and hence is greater than the variance

of the usual intakes. Other parameters of the distribution of mean intakes may also differ from the parameters of the distribution of usual intakes. Because of these problems, using the distribution of the mean of a few days as an estimate of the usual intake distribution can lead to erroneous inferences regarding dietary status.

Nusser, Battese, and Fuller (1990) suggested a measurement error model, where the observed daily intake of an individual is equal to the usual intake for that individual plus a measurement error, and the second and third moment of an individual's measurement errors are modeled as a function of the individual's usual intake. Moment methods are used to estimate the parameters of a specified distribution (e.g., Weibull). Although this approach has the advantage of working with the data in the original scale, it requires several parametric assumptions and is difficult to implement for dietary components that exhibit extreme behavior, such as vitamin A.

A second approach to estimating the usual intake distribution involves transforming the daily intakes so that the transformed values approximately follow a normal distribution. The National Research Council (1986) recommended this approach and suggested a log transformation. As we explain in Section 2, log transformations or simple power transformations do not consistently produce transformed data that are normally distributed.

The problem of estimating the distribution of usual intakes can be formulated as the problem of estimating the distribution of a random variable that is observed subject to measurement error. Mendelsohn and Rice (1982) presented an example of estimation of a density given observations contaminated with normal error. Fan (1991), Stefanski (1990), and Stefanski and Carroll (1990, 1991) studied kernel estimation of an unknown density given observations

S. M. Nusser is Assistant Professor, A. L. Carriquiry is Associate Professor, K. W. Dodd is Research Assistant, and W. A. Fuller is Distinguished Professor, Department of Statistics, Iowa State University, Ames, IA 50011. This research was partly supported by Research Agreement No. 58-3198-9-032 and Cooperative Agreement 58-3198-2-006 between the Agricultural Research Service, U.S. Department of Agriculture, and the Center for Agricultural and Rural Development, Iowa State University. Carriquiry's work was also partially funded by research grant no. N000149610279 from the Office of Naval Research, U.S. Department of Defense. The authors thank Phillip Kott, Patricia Guenther, and the referees for useful comments. They also thank George Battese, who participated in our original research on this topic.

subject to independent additive measurement error with a known distribution. These authors also provided references to earlier work. Stefanski and Bay (1996) described a simulation extrapolation estimation procedure for the cumulative distribution function of observations subject to additive normal errors.

Our approach differs somewhat from the kernel estimators in that we assume that a transformation exists such that both the original observations and the measurement errors are normally distributed. The transformation is a grafted cubic equation fit to a power of the original data. This fitting can be considered a semiparametric version of the Lin and Vonesh (1989) procedure. It is also related to the spline approach for estimating the distribution function (see Wahba 1975 and Wegman 1982). The transformed daily intake data are assumed to follow a measurement error model, and normal theory is used to estimate the parameters of the model. An estimated inverse transformation carries the normal usual intake distribution back to the original scale and defines the distribution of usual intakes. The data set discussed in this article and information on obtaining software to implement our procedure for estimating usual intake distributions have been submitted to STATLIB.

2. APPLICATION TO CSFII DATA

2.1 The CSFII Data

The data for this study are a subset of the data from the 1985 Continuing Survey of Food Intakes by Individuals (CSFII) conducted by the U.S. Department of Agriculture (1987). Daily dietary intakes were collected from women between age 19 and 50 and from the preschool children of the women at approximate 2-month intervals over the period April 1985–March 1986. Twenty-four-hour recall data were collected by personal interview for the first day and by telephone whenever possible for subsequent days. The sample was a multistage stratified area probability sample from the 48 coterminous states and was designed to be self-weighting. Because of the relatively high attrition rate for the 6-day sample, the USDA constructed a 4-day data set for analyses which consisted of the first day of dietary intakes for all individuals who provided at least 4 days of data, plus a random selection of three daily intakes from the remaining 3, 4, or 5 days of available data. Weights were developed to adjust for nonresponse, and the analyses of this article are constructed on the weighted data.

We analyze a subset of the 4-day data set containing dietary intakes for 737 women age 25–50 who were responsible for meal planning within the household and who were not pregnant or lactating during the survey period. Because of the time separation of the observations, we assume the four observations on each individual to be independent observations on that individual. The dietary components calcium, energy, iron, protein, vitamin A, and vitamin C were selected for analysis because of their nutritional importance and because of their varying distributional behaviors.

The report of the National Research Council (1986) provides a review of factors that influence observed daily intakes. Some effects, such as errors in reported food intake

and translation of food intake to nutrient intake, are not estimable from the data of our study. The effect of other factors, such as day of the week, season (month), interview method, and interview sequence, can be investigated.

2.2 Method Overview

We begin with a set of survey responses on daily intakes for n individuals. A subset of the individuals must report more than one daily intake. In our application, four days of intakes are available for each individual. Associated with each individual is a survey weight.

The method for estimating usual intake distributions consists of several steps. First, the intake data are adjusted to remove nuisance effects, such as day of week and interview sequence, which are known to influence daily consumption levels. Because intakes recorded on the first sample day are believed to be the most accurate, the daily intakes for each sample day are adjusted to have a mean and variance equal to that of the first sample day (day 1). Other reference standards, such as the grand mean, can be used. These initial adjustments are described in Section 2.3.

The survey weights for the observations reflect the selection probabilities and are incorporated into the analysis by creating an “equal weight sample” from the original sample (Sec. 2.4). The empirical cumulative distribution function is constructed with the nk weighted observations, where k is the number of observations per individual. Then nk equal weight observations are defined to be the values of the inverse empirical cumulative distribution function associated with the nk equally spaced probability values, $(nk)^{-1}(i - .5)$, $i = 1, 2, \dots, nk$.

The third step involves transforming the adjusted equal weight daily intake data to approximate normality using a combination of a power function and a grafted polynomial function. Some dietary components exhibit sizable deviations from normality when transformed with a simple power function, and the grafted polynomial step provides the flexibility required to transform such components to approximate normality. The semiparametric transformation is described in Section 2.5.

In the fourth step, the parameters of the usual intake distribution are estimated in the transformed scale using a measurement error framework (Sec. 2.6). The measurement error model assumes that transformed daily intakes exhibit heterogeneous within-individual variances, and parameters of the error distribution are estimated.

In the final step of the procedure, the estimated usual intake distribution in the original scale is obtained by applying a back transformation to the normal scale estimated usual intake distribution (Sec. 2.7). The back transformation is estimated by modifying the nonlinear forward transformation so that the back transformation is appropriate for the distribution of individual means.

The steps in the procedure are designed to address the different features of daily dietary intake data, including nuisance effects, survey weights, nonnormality, measurement error, and heterogeneous variances. The sequence in which the issues are addressed and the methods applied is the

product of our attempts to produce a procedure with applicability to a wide range of data possessing these characteristics.

2.3 Initial Adjustments

We begin by adjusting the data for nuisance effects. The adjustment variables will vary with each study. In the case of the 1985 CSFII data, the daily intakes were examined using least squares methods to determine whether day of the week, month, interview mode (telephone or in-person), and interview sequence (first, second, third, or fourth interview) effects were important. Month and interview sequence are confounded to a large degree, because the first interview was conducted at nearly the same point in time for all individuals.

Let W_{0ij} denote the observed intake for the i th individual on the j th day in the interview sequence plus a constant equal to .0001 times the sample mean for the nutrient. This small amount is added to avoid problems in subsequent procedures that depend on the derivative of a power of the data, which can be infinite when evaluated at zero. Consider the sample of n individuals, and let the i th individual have a weight w_i , where $\sum_{i=1}^n w_i = 1$.

Because dietary intake data are often skewed, a power transformation is applied to the data to make the distributions of the observed data more nearly symmetric. To simplify computations, γ is selected by a nonlinear least squares grid search procedure. The original observations W_{0ij} are used to estimate the power γ by minimizing the error sum of squares,

$$\sum_{i=1}^n w_i \sum_{j=1}^k (U_{ij} - \beta_0 - \beta_1 W_{0ij}^\gamma)^2, \quad (1)$$

over a grid of values of γ , where U_{ij} is the normal score for the ij th observation and β_0 and β_1 are estimated for each value of γ . The normal scores are computed as

$$U_{ij} = \Phi^{-1} \left[\left(\frac{s_{ij} - 3}{8} \right) / \left(\frac{nk + 1}{4} \right) \right], \quad (2)$$

where Φ is the standard normal distribution function and s_{ij} is the rank of the ij th observation. The grid of values for γ is $[1, (1.5)^{-1}, (2.0)^{-1}, \dots, (10)^{-1}, \log]$, where \log denotes the natural logarithm and corresponds to $\gamma = 0$.

Once the power has been selected, a model containing day of the week, interview mode, and interview sequence as additive classification variables is fit by weighted least squares to the power-transformed observations, W_{0ij}^γ , where the weights in the regression are the sampling weights. Interview mode is not significant for any dietary component. Day-of-week effects are significant for energy ($p < .001$) and protein ($p < .05$) intakes, primarily because of higher consumption on weekends for both dietary components. Sequence effects (confounded with month effects) are significant at the $\alpha = .001$ level for calcium, energy, iron, and protein intakes and are principally attributable to higher in-

take levels on the first interview day versus the other 3 days.

Because of these results, data were adjusted for weekday and interview sequence effects. Let $Z_{0ij} = W_{0ij}^\gamma$ represent the power-transformed observed intake for the i th individual on the j th day. The ij th observation adjusted for weekday and interview sequence effects is $Z_{ij} = \hat{Z}_{0ij}^{-1} \bar{Z}_{0.1} Z_{0ij}$, where $\bar{Z}_{0.1}$ is the mean of the power-transformed observed intakes for the first interview day and \hat{Z}_{0ij} is the predicted intake from the regression for the i th individual on the j th day. The ratio adjustment is used to reduce the probability that adjusted intake values are nonnegative. Should negative adjusted intakes occur, they are set equal to zero. The data are adjusted to the mean of the first interview day (rather than the grand mean), because the data are believed to be more accurate on the first interview day than on subsequent days.

It is well established that the characteristics of responses in a repeated survey are a function of the time in sample at which a respondent is observed (see, e.g., Bailar 1975). Our initial regression adjustment modifies the data so that there is no sequence effect in the mean of the intake distributions for the different days. Because of the possibility of other higher-order time-in-sample effects, we standardized the sample variance of transformed observations for the second, third, and fourth times in sample to the sample variance observed on the first day. The adjusted observations in the original scale are defined by

$$Y_{ij}^* = [\hat{\mu}_{.1} + S_{.j}^{-1} S_{.1} (Z_{ij} - \hat{\mu}_{.j})]^{1/\gamma}, \quad (3)$$

where $i = 1, 2, \dots, n$ individuals, $j = 1, 2, \dots, k$ days, $S_{.j}^2 = (n-1)^{-1} \sum_{i=1}^n (Z_{ij} - \hat{\mu}_{.j})^2$, and $\hat{\mu}_{.j} = n^{-1} \sum_{i=1}^n Z_{ij}$. For a very few observations (fewer than four for every component), the transformation is modified near zero to guarantee nonnegative adjusted data.

The among- and within-individual standard deviations for the adjusted intakes in original units as defined in (3) are presented in Table 1. These statistics indicate that there is considerable within-individual variability relative to among-individual variability. The ratios of within- to among-individual variances are similar to those for comparable dietary components reported by the National Research Council (1986). Vitamin A is unusual in that there is one very large observation and a few other large observations that are responsible for the very large within-individual variance. Table 1 also contains the estimator of skewness, where skewness is defined as the third central moment divided by the cube of the standard deviation. The skewness coefficient indicates that for most dietary components, an assumption of normality is unreasonable. In addition, analyses not shown here indicate that within-individual standard deviations are positively correlated with individual means.

2.4 Incorporating Survey Weights

Our estimation scheme is designed to handle samples with unequal weights. To apply classical equal-weight methods for the estimation of the components of variance

Table 1. Sample Moments for Adjusted Observed Intakes Y_{ij}^* in (3)

Dietary component	Mean	Among-individual std. dev.	Within-individual std. dev.	Ratio of within- to among-individual variances	Skewness
Calcium (mg)	622.3	253.9	319.1	1.58	1.35
Energy (kcal)	1,683.4	450.4	585.1	1.69	1.10
Iron (100 mg)	1,105.3	294.9	482.0	2.67	1.76
Protein (10 g)	668.8	157.5	270.5	2.95	1.38
Vitamin A ($\mu\text{g}/\text{RE}$)	801.0	570.9	1,401.0	6.02	11.75
Vitamin C (10 mg)	792.5	408.6	625.6	2.34	1.87

NOTE: Data are adjusted for day-of-week and interview sequence effects.

model, we use the empirical cumulative distribution function to create an equal-weight sample from the adjusted unequal-weight sample. The first step in creating the equal-weight sample is to construct an empirical cumulative distribution function from the nk observations, defined by

$$\hat{F}_Y(a) = \sum_{i=1}^n w_i \sum_{j=1}^k I_{Y_{ij}}(a),$$

where $I_{Y_{ij}}(a)$ is the indicator function with

$$I_{Y_{ij}}(a) = \begin{cases} 1 & \text{if } Y_{ij}^* \leq a, \\ 0 & \text{otherwise.} \end{cases}$$

A continuous function, denoted by $\tilde{F}_Y(a)$, is created by connecting the midpoints of the rises in the steps of $\hat{F}_Y(a)$. This function is used to define nk observations of an equal-weight sample that gives nearly the same distribution function as that of the adjusted data. The equal-weight observations are defined by $Y_{ij} = \tilde{F}_Y^{-1}[(nk)^{-1}(s_{ij} - .5)]$ for $i = 1, 2, \dots, n$ and $j = 1, \dots, k$, where s_{ij} is the rank of the Y_{ij}^* . These adjusted, equal-weight intakes Y_{ij} are hereafter called daily intakes.

2.5 Semiparametric Transformation to Normality

The first step to transforming the daily intakes to normality is to calculate normal scores U_{ij} , as defined in (2) for the Y_{ij} . The pairs (U_{ij}, Y_{ij}) are used to estimate a semiparametric function that transforms the daily intakes into approximately normal variables. The transformation function is fit to the data in two phases. First, a power is determined that produces observations that are close to normally distributed by minimizing (1), where $w_i \equiv 1$ and Y_{ij} replaces

W_{0ij} . Let the selected power be denoted by α . The inverses of the powers of the first transformation step for the CSFII data are given in the first column of Table 2.

In the next phase of the normality transformation, a grafted cubic polynomial is fit to the (U_{ij}, Y_{ij}^α) pairs, minimizing deviations in the Y direction. Let the join points for the polynomial be B_1, B_2, \dots, B_p . The values of B_1 and B_p are chosen such that two data points are outside of each of the outside join points. For our data, B_1 and B_p are -3.26 and 3.26 . The values B_2, \dots, B_{p-1} are defined such that the intervals $(B_i, B_{i+1}), i = 1, 2, \dots, p-1$ are of equal length. The function is constructed to be linear for $U_{ij} < B_1$, linear for $U_{ij} > B_p$, and cubic in the intervals $(B_i, B_{i+1}), i = 1, 2, \dots, p-1$, with continuous first and second derivatives at the join points. (See, e.g., Fuller 1976, p. 393, for a description of the function.) The fitted grafted polynomial function is also constrained to be monotone increasing. At least three join points are included in the model for each component. The number of parameters, p , in the grafted polynomial model is equal to the number of join points.

The number of join points for the grafted cubic is chosen to be the minimum number of join points (up to 12) required to make the value of the Anderson–Darling test statistic for normality less than .58 when applied to the data transformed by the semiparametric function. The Anderson–Darling test was chosen as a “goodness-of-fit” criterion to measure the distance between the distribution of the transformed observations and the normal distribution (see Anderson and Darling 1952 and Stephens 1974). The value of .58 is approximately the 15% point of the distribution. Using the 15% point as the cutoff is analogous to adding a variable to a regression when the F statistic exceeds 2 and is also analogous to using the Akaike information criterion

Table 2. Statistics for the Semiparametric Transformation to Normality

Dietary component	Inverse of power	Anderson–Darling for fitted values ^a	Number of join points	t for heterogeneous variances ^b	t for linear effect ^b
Calcium	3.5	.28	3	2.72	−1.80
Energy	2.0	.47	4	2.46	−.02
Iron	2.5	.40	5	2.22	−.50
Protein	2.0	.21	6	2.01	.57
Vitamin A	5.5	.36	11	3.28	−.96
Vitamin C	3.5	.34	8	2.06	−1.47

^a Reject at the 15% level if the Anderson–Darling statistic is greater than .58.^b Reject the null hypothesis of zero slope at the 5% level if $|t| > 1.96$.

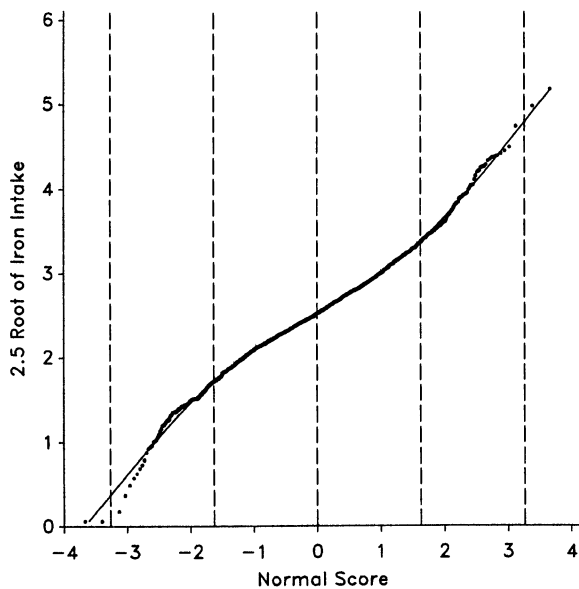


Figure 1. Plot of Grafted Polynomial for Iron. The plot of power-transformed daily intakes versus normal scores is represented by points. The smooth line is the fitted grafted polynomial. Vertical dashed lines designate join points of the grafted polynomial.

(AIC) (see Sawa 1978 and Sawa and Hiromatsu 1973). The Anderson–Darling statistic computed for the daily intake data transformed with the grafted polynomial is given in the second column of Table 2. The number of join points is given in the third column of Table 2.

Figure 1 contains a plot of the $(2.5)^{-1}$ power of the iron daily intakes against the normal scores. The S-shaped plot indicates that a simple power transformation is not adequate to transform the plot into a straight line. Note that there are 2,948 observations in the figure, with 2 to the left of the left join point and 2 to the right of the right join point. Because the function is constrained to be linear beyond the outside join points and because the function has a continuous second derivative, it is possible that the fitted function will not pass through the extreme observations.

Although the total set of observations is transformed to normality, it does not follow automatically that the individual means are normally distributed. The normality of individual means is of interest, because their distribution is more closely related to the distribution of usual intakes than that of the original observations. Because of this, as an additional check on the transformation, the Anderson–Darling statistic was computed for the individual means of the transformed daily intakes. In no case was the statistic significant at the 10% level.

To check the hypothesis that the within-individual variances calculated from the transformed daily intake data are constant over individuals, let

$$A_i = (k-1)^{-1} \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2$$

and

$$\hat{M}_4 = 3\bar{A}^{-2}n^{-1} \sum_{i=1}^n [1 + 2(k-1)^{-1}]^{-1} A_i^2, \quad (4)$$

where X_{ij} is the transformed value for individual i on day j , $\bar{X}_i = k^{-1} \sum_{j=1}^k X_{ij}$, $\bar{A} = n^{-1} \sum_{i=1}^n A_i$, and $k = 4$ is the number of observations per individual. If the transformed observations are normally distributed with homogeneous variances and four observations per person, \hat{M}_4 estimates 3, the fourth moment of the standard normal distribution. The approximate variance is

$$V(\hat{M}_4) = 9n^{-1}(k-1)^{-4}[1 + 2(k-1)^{-1}]^{-2} \times \left\{ \left[2^4 \Gamma\left(\frac{k-1}{2}\right)^{-1} \Gamma\left(\frac{k-1}{2} + 4\right) \right] - [2(k-1) + (k-1)^2]^2 \right\},$$

which is equal to .039077 for $n = 737$ and $k = 4$. The values of the test statistic

$$[V(\hat{M}_4)]^{-1/2}(\hat{M}_4 - 3)$$

calculated using the transformed daily data is given in Table 2 under the heading “ t for heterogeneous variances.” This ratio is greater than 1.96 for all nutrients analyzed, indicating that the within-individual variances vary across individuals.

To investigate the hypothesis that the heterogeneity of within-individual variances in the transformed space is due to a relationship between within-individual standard deviations and individual means, the model $A_i^{1/2} = \beta_0 + \beta_1 \bar{X}_i$ was fit using least squares. The t statistics for testing the hypothesis that $\beta_1 = 0$ are presented in column 4 of Table 2. The statistic for calcium is -1.80 , whereas the remaining statistics are less than 1.5 in absolute value. When within-individual standard deviations are plotted against individual means, no obvious patterns are revealed. Therefore, it was decided to complete the analysis for all nutrients under the assumption that the variances are not related to the means.

2.6 Estimating the Usual Intake Distribution in the Normal Scale

A measurement error model is used for estimating the distribution of usual intakes in the normal scale. Let

$$X_{ij} = x_i + u_{ij}, \quad (5)$$

where $x_i \sim NI(\mu_x, \sigma_x^2)$, $u_{ij} \sim N(0, \sigma_{ui}^2)$, $\sigma_{ui}^2 \sim (\mu_A, \sigma_A^2)$, x_i is the unobservable usual intake value for individual i in the

Table 3. Estimated Moments for Normal-Scale Daily Intakes, X_{ij}

Dietary component	Among-individual variance $\hat{\sigma}_x^2$	Average within-individual variance $\hat{\mu}_A$	Within-to among-ratio $\frac{\hat{\mu}_A}{\hat{\sigma}_x^2}$	Variance of individual variances $\hat{\sigma}_A^2$
Calcium	.367	.635	1.73	.072
Energy	.378	.626	1.66	.064
Iron	.318	.685	2.15	.069
Protein	.273	.728	2.67	.070
Vitamin A	.261	.742	2.84	.119
Vitamin C	.320	.684	2.14	.064

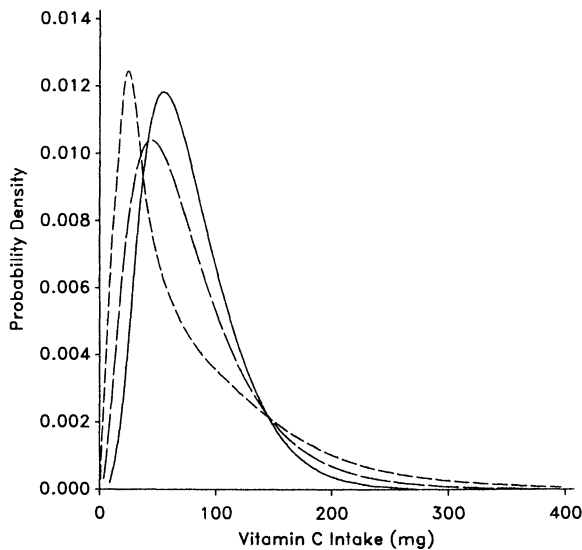


Figure 2. Estimated Densities of Usual Intakes (Usual; Solid Line), 4-Day Individual Means (Mean; Long-Dashed Line), and 1-Day Intakes for Vitamin C (Daily; Short-Dashed Line).

normal scale, u_{ij} is the unobservable measurement error for individual i on day j in the normal scale, the u_{ij} are independent given i , and x_i and u_{lj} are independent for all i, l , and j .

On the basis of the empirical analyses presented in Table 2, we permit heterogeneous within-individual variances. The errors u_{ij} represent variation of two kinds: the day-to-day variability in the true amounts eaten by individual i and the difference between the true amount eaten and the amount reported for an individual. It is believed that the day-to-day variance for an individual is much larger than the variance of the reporting error. The transformed daily intakes X_{ij} have $\mu_X \doteq 0$ and $\sigma_X^2 \doteq 1$. Under model (5), the conditional distribution of X_{ij} , given (x_i, σ_{ui}^2) is $N(x_i, \sigma_{ui}^2)$. However, the unconditional distribution is not normal if $\sigma_A^2 > 0$. We conduct our analysis under the operational assumption that the initial transformation produces x_i and u_{ij} satisfying (5). Under (5), the individual means, $\bar{X}_i = k^{-1} \sum_{j=1}^k X_{ij}$, are independent $(0, \sigma_{\bar{X}}^2)$ random variables, where $\sigma_{\bar{X}}^2 = \sigma_x^2 + k^{-1}\mu_A$. For our purposes, it is not necessary to specify a form of the distribution of the individual error variances, because we will only use the variance of the distribution of variances.

Table 4. Estimated Moments for Usual Intakes in the Original Scale, y_i

Dietary component	Mean	Standard deviation	Skewness
Calcium (mg)	622.4	240.0	.84
Energy (kcal)	1,684.5	443.9	.54
Iron (100 mg)	1,107.4	305.9	.75
Protein (10 g)	670.0	160.1	.52
Vitamin A ($\mu\text{g RE}$)	822.0	536.3	2.64
Vitamin C (10 mg)	791.2	397.2	1.08

Estimators for the moments are

$$\hat{\mu}_x = n^{-1} \sum_{i=1}^n \bar{X}_i,$$

$$\hat{\sigma}_{\bar{X}}^2 = (n-1)^{-1} \sum_{i=1}^n (\bar{X}_i - \hat{\mu}_x)^2,$$

$$\hat{\mu}_A = [n(k-1)]^{-1} \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2,$$

$$\hat{\sigma}_x^2 = \hat{\sigma}_{\bar{X}}^2 - k^{-1} \hat{\mu}_A,$$

and

$$\hat{\sigma}_A^2 = n^{-1} [1 + 2(k-1)^{-1}] \sum_{i=1}^n A_i^2 - \bar{A}^2,$$

where A_i and \bar{A} are defined in (4). The within- and among-individual variances for the transformed data are given in Table 3. In all cases the sum of the within-individual and among-individual variances is close to 1, because the transformed data have mean 0 and variance 1. The average of the within-individual variances exceeds the among-individual variance for all dietary components. The ratio of within- to among-individual variance is smallest for energy with a value of 1.66 and is largest for vitamin A with a ratio of 2.84. The ratios of within- to among-individual variance of Table 3 are similar to the corresponding ratios computed from the standard deviations in original scale of Table 1, with the exception of vitamin A. In the original scale, the vitamin A data are skewed, the individual standard deviations are positively correlated with the individual means, and a few very large observations made a large contribution to the within-individual variance in the original scale.

The last column of Table 3 contains an estimate of the variance of the individual variances, denoted by $\hat{\sigma}_A^2$. The coefficients of variation of the individual variances are 36%–46%.

2.7 Estimated Usual Intake Distribution in the Original Scale

Under model (5), the conditional distribution of observed daily intakes in the normal scale for all individuals with a particular usual intake, say \ddot{x}_i , is the average of all normal distributions with common mean \ddot{x}_i and variance A , where $A \sim (\mu_A, \sigma_A^2)$. Thus daily intake in normal scale is the sum of \ddot{x}_i and u , where

$$E\{(u, u^2, u^4) | x = \ddot{x}_i\} = (0, \mu_A, 3\mu_A^2 + 3\sigma_A^2)$$

and the distribution of u is symmetric about 0.

Let \ddot{y}_i denote the usual intake in original scale for all individuals with usual normal intake \ddot{x}_i , and let g denote the transformation taking the adjusted observed intakes Y to normality. Then

$$\ddot{y}_i = E\{Y | x = \ddot{x}_i\} = E\{g^{-1}(x + u) | x = \ddot{x}_i\} = h(\ddot{x}_i).$$

Table 5. Sample Moments for Individual 4-Day Means in the Original Scale, Calculated From Daily Intakes, Y_{ij}^*

Dietary component	Mean	Standard deviation	Skewness
Calcium (mg)	622.3	295.3	1.18
Energy (kcal)	1,683.4	532.9	.70
Iron (100 mg)	1,105.3	387.7	1.17
Protein (10 g)	668.8	208.4	.89
Vitamin A (μg RE)	800.8	864.9	7.17
Vitamin C (10 mg)	792.4	515.6	1.24

NOTE: Data are adjusted for day-of-week and interview sequence effects.

The transformation h is estimated by approximating the conditional expectation of Y at a set of values of \tilde{x} and then fitting a grafted polynomial to the (\tilde{y}, \tilde{x}) pairs. The set of \tilde{x} values is a set of 400 values such that the first five moments of the points match the first five moments of a $N(0, \hat{\sigma}_x^2)$ distribution. At each value of \tilde{x} , we use a nine-point approximation to the distribution of u . The distribution of u has mean zero and a variance with estimated mean $\hat{\sigma}_u^2$ and estimated variance of variance equal to $\hat{\mu}_A$. Nine points, c_l , and nine weights, w_l , where $\sum w_l = 1$, are constructed such that the first five moments of the discrete nine-point distribution match the first five estimated moments of the conditional distribution of $\tilde{x} + u$ conditional on \tilde{x} . For each of the 400 values of \tilde{x} , the usual intake in the original scale is approximated by $\tilde{y}_i = \sum_{l=-4}^4 w_l g^{-1}(\tilde{x}_i + c_l)$, where \tilde{x}_i is the i th value in normal scale and c_l and w_l ($l = -4, -3, \dots, 4$) are the values and weights for the nine-point approximation to the distribution of u . The 400 \tilde{y} values provide a 400-point estimator of the usual intake distribution. A grafted cubic created from the pairs $(\tilde{y}_i, \tilde{x}_i)$, denoted by \hat{h} , is an estimator of the transformation of the normal \tilde{x} into the usual intake \tilde{y} in the original scale.

Densities for the dietary components were constructed by multiplying the derivative of $\hat{h}^{-1}(y)$ by the normal ordinate for the usual intake density of the component in the normal scale. The estimated density of usual intakes for vitamin C is the solid line in Figure 2. Also in the figure is the estimated density for daily intakes, identified by the short dashed line, and the estimated density of the 4-day means, identified by the long dashed line. The estimated density

for 4-day means was approximated by applying the same smoothing algorithm used to estimate the distribution function of daily intakes to the individual means. The skewness in the density of the mean declines as the number of daily intakes in the mean increases.

Table 4 contains the mean, variance, and skewness coefficient for the estimated usual intake distributions calculated from the 400-point approximation. The estimated means of the usual intakes are very close to the means for the daily intakes adjusted for nuisance effects and time-in-sample effects (Y_{ij}^*) presented in Table 1. Also, the estimated standard deviations of usual intakes are close to the among-individual standard deviations of Table 1 for calcium, energy, iron, and protein. This is to be expected, because the estimates of Table 1 are the sample moment estimators of the same quantities. The estimated standard deviations of usual intakes for the two vitamins differ considerably from the direct moment estimators of Table 1. As previously mentioned, the original distributions for the vitamin daily intakes are very skewed and, hence the original sample moments are heavily dependent on a few large observations. The effect on the large observations is reduced for estimators constructed using our procedure.

Comparison of the sample moments for usual intakes in Table 4 with the estimated moments for individual means (Table 5) indicates that the distribution of 4-day means is a poor estimate of the usual intake distribution. For all dietary components, the standard deviation and skewness coefficient are larger for the mean distribution than for the estimated usual intake distribution.

Table 6 contains estimated percentiles for the usual intakes of six dietary components. The percentiles were computed with the estimated transformation function using the percentiles of the estimated distribution of usual intakes in normal scale. For example, the estimated mean and variance of vitamin C usual intakes in the normal scale are zero and .320. Therefore, the estimated 95% point in the normal scale is $\hat{\sigma}_x \Phi^{-1}(.95) = .566 \times 1.645 = .931$. Using the estimated h transformation, the 95% point of the usual intake distribution in original scale is 154.7 mg.

A balanced repeated replication method was used to estimate the standard deviations of the estimated percentiles

Table 6. Estimated Percentiles for Usual Intake Distributions in the Original Scale

Component	Percentile						
	.01	.05	.10	.50	.90	.95	.99
Calcium (mg)	208 (14)	292 (13)	345 (13)	590 (13)	942 (32)	1,065 (41)	1,326 (64)
Energy (kcal)	796 (38)	1,023 (34)	1,151 (33)	1,648 (36)	2,262 (56)	2,469 (66)	2,911 (95)
Iron (100 mg)	527 (27)	670 (26)	751 (26)	1,074 (30)	1,504 (52)	1,657 (63)	1,998 (91)
Protein (10 g)	349 (18)	431 (17)	477 (17)	658 (15)	878 (21)	951 (26)	1,108 (39)
Vitamin A (μg RE)	218 (16)	312 (18)	372 (19)	671 (35)	1,449 (134)	1,837 (200)	2,876 (442)
Vitamin C (10 mg)	184 (16)	287 (19)	356 (21)	717 (33)	1,328 (61)	1,547 (72)	2,014 (97)

NOTE: Values in parentheses are estimated standard errors.

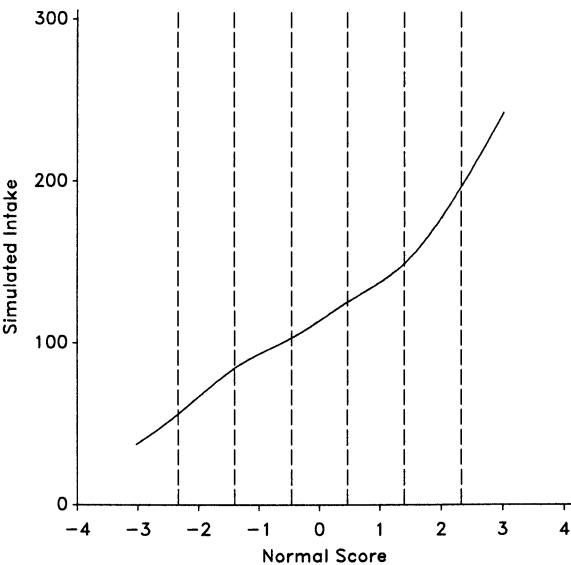


Figure 3. Plot of Function Used to Generate Original Scale Daily Intakes From Normal Daily Intakes for Simulation.

(see, e.g., Fay 1989 and Wolter 1985). The sample is a stratified sample with two primary sampling units per stratum. Some strata were combined to create a sample of 48 strata, each with two primary sampling units. Sixteen replicates were created based on orthogonal contrasts. Each replicate contains one-half of the data. All operations, including the power and grafted polynomial estimation, were carried out for each of the replicate samples. The estimated standard errors of the estimated percentiles, given below the estimates in Table 6, are the square roots of $\hat{V}\{\hat{\theta}_0\} = (16)^{-1} \sum_{i=1}^{16} (\hat{\theta}_i - \hat{\theta}_0)^2$, where $\hat{\theta}_i$ is the estimated percentile for the i th replicate and $\hat{\theta}_0$ is the estimate for the original sample.

3. MONTE CARLO STUDY

We conducted a Monte Carlo study to evaluate the performance of the estimation procedure described in Section 2 and to compare our method with two other procedures. The first alternative procedure for estimating the distribution of usual intakes is comprised of the following steps:

- 1. Power transform the daily intakes, where the selected power is chosen to minimize the Anderson–Darling statistic.
- 2. Compute the mean daily intake for each individual using the transformed data.
- 3. Shrink the individual means of the transformed data for individual i as follows: $\tilde{x}_i = \hat{\mu}_x + \hat{\sigma}_x^{-1} \hat{\sigma}_x (\bar{X}_i - \hat{\mu}_x)$, where $\hat{\mu}_x$ is the mean of the transformed observations, $\hat{\sigma}_x^2$ is the estimated variance of the transformed means, and $\hat{\sigma}_x^2$ is the estimator of the among-individual variance. The shrunken means have the mean and variance of the usual distribution in the transformed scale.
- 4. Back-transform the shrunken means to the original scale using a Taylor series approximation to adjust for bias when applying the inverse nonlinear transformation to usual intakes.

- 5. Estimate the cumulative distribution function of usual intakes from the back-transformed shrunken means by the empirical distribution function.

This procedure is an extension of the suggestions of the National Research Council (1986). Because the primary difference between the procedure described above and our procedure of Section 2 is in the transformation of step 1, we call the outlined procedure the *best power* procedure.

The second alternative procedure for estimating usual intake distributions is based on the smoothed empirical distribution of individual mean intakes. This method has been used in the past by practitioners and is expected to do poorly in the tails, because of the presence of within-individual variation in the distribution.

In the simulation, a true usual intake distribution was generated that displays distributional characteristics similar to those of protein. Protein is in the center of the components studied with respect to skewness and with respect to number of join points. For each of 1,000 samples, an observation Y_{ij} for the j th day ($j = 1, 2$) on the i th individual ($i = 1, \dots, 700$) was generated as follows:

- Draw x_i , the individual’s usual intake in normal scale from a $N(0, .36)$ distribution.
- Draw σ_{ui}^2 , the measurement error variance, from a uniform distribution on the values .32, .50, .64, 1.1. The measurement error variance distribution has mean .64, and variance .0834.

Table 7. Estimates of Selected Percentiles of the Usual Intake Distribution Using Three Estimation Methods, Averaged Over 1,000 Simulations

Percentile	True	Estimation method		
		Spline	Best power	2-day mean
.01	81.85	81.85 (.076) [2.402]	79.31 (.103) [4.142]	67.56 (.108) [14.690]
.05	91.86	91.88 (.052) [1.638]	92.03 (.062) [1.952]	83.40 (.059) [8.659]
.10	96.99	96.91 (.043) [1.355]	97.97 (.048) [1.805]	91.02 (.045) [6.138]
.25	105.44	105.27 (.033) [1.057]	106.18 (.036) [1.351]	101.99 (.035) [3.625]
.50	115.03	115.03 (.029) [.931]	115.17 (.032) [1.011]	114.49 (.033) [1.176]
.75	125.23	125.49 (.039) [1.255]	124.61 (.039) [1.390]	127.77 (.040) [2.831]
.90	135.42	135.70 (.058) [1.843]	134.13 (.060) [2.290]	141.62 (.065) [6.527]
.95	142.23	142.47 (.075) [2.395]	141.50 (.083) [2.732]	152.69 (.095) [10.871]
.99	157.00	157.48 (.128) [4.080]	159.27 (.157) [5.467]	180.05 (.208) [23.956]

NOTE: Values in parentheses are estimated standard errors for the Monte Carlo mean percentiles. Values in brackets are estimated RMSEs.

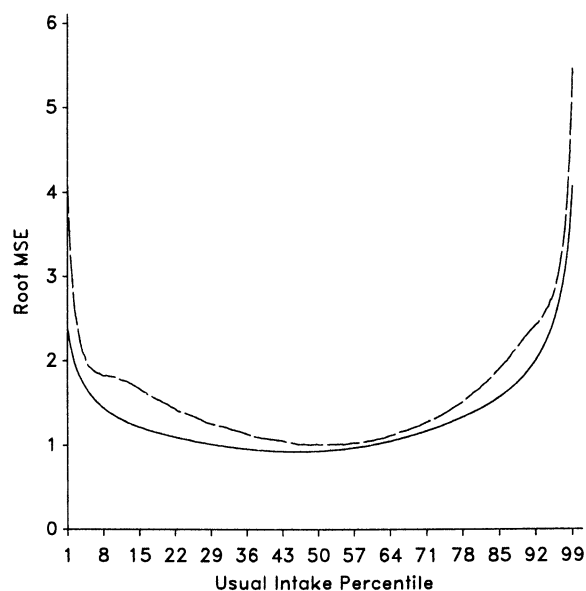


Figure 4. Average Estimated Root Mean Squared Error in Estimated Percentiles From 1,000 Simulation Runs, for the Proposed (Spline; Solid Line) and Best Power (BP; Dashed Line) Estimation Methods.

- Draw the measurement error u_{ij} from a normal distribution with mean zero and variance σ_{ui}^2 , for $j = 1, 2$, and form $X_{ij} = x_i + u_{ij}$, where X_{ij} is the daily intake in normal scale. If X_{ij} falls below -6.97 , then X_{ij} is set equal to -6.97 .

Let $Y_{ij} = L_{ij}^{2.5}$ be the daily intake in the original scale, where L_{ij} is a grafted cubic function of X_{ij} . The definition for L_{ij} is such that no power of the generated intakes is normally distributed. The function relating Y_{ij} and X_{ij} is presented in Figure 3.

We computed 291 percentiles of the estimated usual intake distributions using the three procedures, and averaged these over the 1,000 samples. The set of percentiles is de-

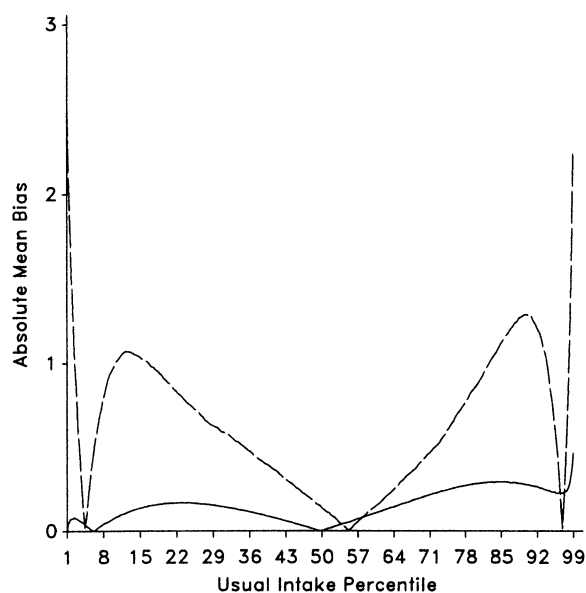


Figure 5. Absolute Average Estimated Bias From 1,000 Simulation Runs, for the Proposed (Spline; Solid Line) and Best Power (BP; Dashed Line) Estimation Methods.

fined by the 41 percentiles .01, .025 to .975 by .025, .99, plus percentiles corresponding to 250 equally spaced probabilities. The estimated percentiles were compared to the percentiles of the true usual intake distribution, generated by numerical integration. Results for selected percentiles are shown in Table 7. The estimated root mean squared error (RMSE) and the absolute value of the average estimated bias are plotted against each of the 291 percentiles for the three estimation methods in Figures 4 and 5. The RMSE and, to some degree, the bias are larger in the tails than in the center of the distribution for all procedures. Although all three methods display some bias, the method proposed in Section 2, called "spline" in the table, produces estimates of percentiles that are nearly always less biased than the other two methods. The spline method generally has smaller standard errors than the best power procedure, especially in the tails, and is uniformly superior to the best power procedure with respect to MSE for all 291 percentiles calculated in the simulation. As expected, the method proposed in Section 2 provides less biased and less variable estimates than estimates based on individual means of the 2 days. The distribution estimated using individual means is comparable to the other procedures only for percentiles near the mean of the usual intake distribution.

To investigate the performance of the balanced repeated replication variance estimation procedure, the 700 individuals in each sample were randomly assigned to 32 approximately equal sized clusters in 16 strata containing 2 clusters each. Variances based on the 16 balanced replicates were computed for the quantiles and confidence intervals calculated. The coverage of the nominal 95% intervals were .959, .964, .967, .963, .958, .943, .941, .943, and .948 for the .01, .05, .10, .25, .50, .75, .90, .95, and .99 percentiles. As these coverages suggest, the replication variances averaged somewhat larger than the Monte Carlo variances for the first three-fourths of the quantiles and somewhat less than the Monte Carlo variance for the larger quantiles. The general performance of the replication intervals was judged satisfactory, although the coverage of some intervals differs significantly (at the 5% level) from the 95% nominal level.

4. COMMENTS

We have presented a method for estimating distributions of usual intakes based on daily intakes of dietary components consumed almost every day. This method is applicable more broadly to settings where the distribution of nonnormal unobservable means is of interest, and the observed data are repeated measurements on a sample unit that measure the sample unit's mean with considerable error. An example of such a problem is determining an individual's average blood pressure using multiple measurements.

This approach is being extended to estimating usual intake distributions for dietary components that are not consumed daily, such as individual foods. The daily intake data for such components contain numerous zeros from individuals who never consume the food and from those who did not eat the food on the sample days. Nusser et al. (1997) proposed methods for estimating the usual intake distribu-

tion for foods when usual intakes are uncorrelated with the frequency of consumption.

In addition, distributions of usual intakes of ratios of dietary components, such as percent calories from fat, are of interest. When both the numerator and denominator in the ratio are observed with error, the methods presented in this article are not directly applicable. A modification to the approach that is suitable for estimating distributions of usual ratio intakes was presented by Carriquiry, Fuller, Goyeneche, and Dodd (1995).

[Received November 1991. Revised May 1996.]

REFERENCES

- Anderson, T. W., and Darling, D. A. (1952), "Asymptotic Theory of Certain 'Goodness-of-Fit' Criteria Based on Stochastic Processes," *Annals of Mathematical Statistics*, 23, 193–212.
- Bailar, B. A. (1975), "The Effects of Rotation Group Bias on Estimates From Panel Surveys," *Journal of the American Statistical Association*, 70, 23–29.
- Carriquiry, A. L., Fuller, W. A., Goyeneche, J. J., and Dodd, K. W. (1995), "Estimation of the Usual Intake Distributions of Ratios of Dietary Components," Staff Report 95-SR 79, Iowa State University, Center for Agricultural and Rural Development.
- Fan, J. (1991), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19, 1257–1272.
- Fay, R. E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 212–217.
- Fuller, W. A. (1976), *Introduction to Statistical Time Series*, New York: Wiley.
- Lin, L. I.-K., and Vonesh, E. F. (1989), "An Empirical Nonlinear Data-Fitting Approach for Transforming Data to Normality," *The American Statistician*, 43, 237–243.
- Mendelsohn, J., and Rice, J. (1982), "Deconvolution of Microfluorometric Histograms With *B*-Splines," *Journal of the American Statistical Association*, 77, 748–753.
- National Research Council (1986), *Nutrient Adequacy*, Washington, DC: National Academy Press.
- Nusser, S. M., Battese, G. E., and Fuller, W. A. (1990), "Method-of-Moments Estimation of Usual Nutrient Intakes Distributions," Working Paper 90-WP52, Iowa State University, Center for Agricultural and Rural Development.
- Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1997), "Estimating Usual Dietary Intake Distributions: Adjusting for Measurement Error and Nonnormality in 24-Hour Food Intake Data," in *Survey Measurement and Process Quality*, eds. L. Lyberg, M. Collins, E. DeLeeuw, C. Dippo, W. Schwartz and D. Trewn, New York: Wiley, pp. 689–709.
- Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46, 1273–1291.
- Stefanski, L. A. (1990), "Rates of Convergence of Some Estimators in a Class of Deconvolution Problems," *Statistics and Probability Letters*, 9, 229–235.
- Stefanski, L. A., and Bay, J. M. (1996), "Simulation Extrapolation Deconvolution of Finite Population Cumulative Distribution Function Estimators," *Biometrika*, 83, 407–417.
- Stefanski, L. A., and Carroll, R. J. (1990), "Deconvoluting Kernel Density Estimators," *Statistics*, 21, 169–184.
- (1991), "Deconvolution-Based Score Tests in Measurement Error Models," *The Annals of Statistics*, 19, 249–259.
- Stephens, M. A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730–737.
- U.S. Department of Agriculture, Human Nutrition Information Service (1987), *Continuing Survey of Food Intakes by Individuals, Women 19–50 Years and Their Children 1–5 years, 4 days, 1985*. CSFII Report No 85-4, Washington, DC: U.S. Government Printing Office.
- Wahba, G. (1975), "Interpolating Spline Methods for Density Estimation, I. Equal Spaced Knots," *The Annals of Statistics*, 3, 30–48.
- Wegman, E. J. (1982), "Density Estimation," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, New York: Wiley.
- Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.